(12) UK Patent Application (19) GB (11) 2 339 643 (13) A

(43) Date of A Publication 02.02.2000

(21) Application No 9810573.7

(22) Date of Filing 18.05.1998

(51) INT CL[7]
H04M 3/523

(52) UK CL (Edition R)
H4K KF50E KF50X

(56) Documents Cited
GB 2315384 A WO 96/22650 A1 WO 92/07318 A1
US 5335268 A

(58) Field of Search
UK CL (Edition Q) H4K KF50A KF50E KF50X
INT CL[6] H04M

(71) Applicant(s)
Callscan Limited
(Incorporated in the United Kingdom)
Priestley Wharf, 20 Holt Street, BIRMINGHAM, b7 4bz, United Kingdom

(72) Inventor(s)
Anthony Scragg
Roger Huffadine

(74) Agent and/or Address for Service
Chris J Tillbrook & Co
1 Mill Street, WARWICK, CV34 4HB, United Kingdom

(54) Abstract Title
Call centre management

(57) A (call centre) telephone call (handling) management system (64) is configured to optimise allocation of appropriately skilled agent operators (50) for call handling, in providing a predetermined 'level of service'; by monitoring (36) call density and deploying knowledge (45, 46) of agent resource skill profiles, availability and attendant budget limitations and applying those factors to a modeling algorithm - allowing call allocation across agent skill groups, or even on a one-to-one basis with individual agents.

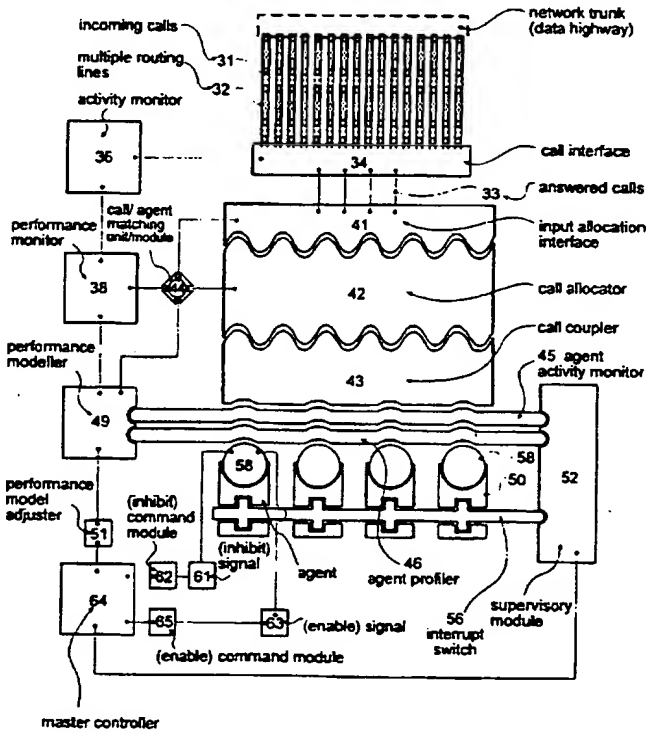

Figure 8

At least one drawing originally filed was informal and the print reproduced here is taken from a later filed formal copy.

This print takes account of replacement documents submitted after the date of filing to enable the application to comply with the formal requirements of the Patents Rules 1995

GB 2 339 643 A

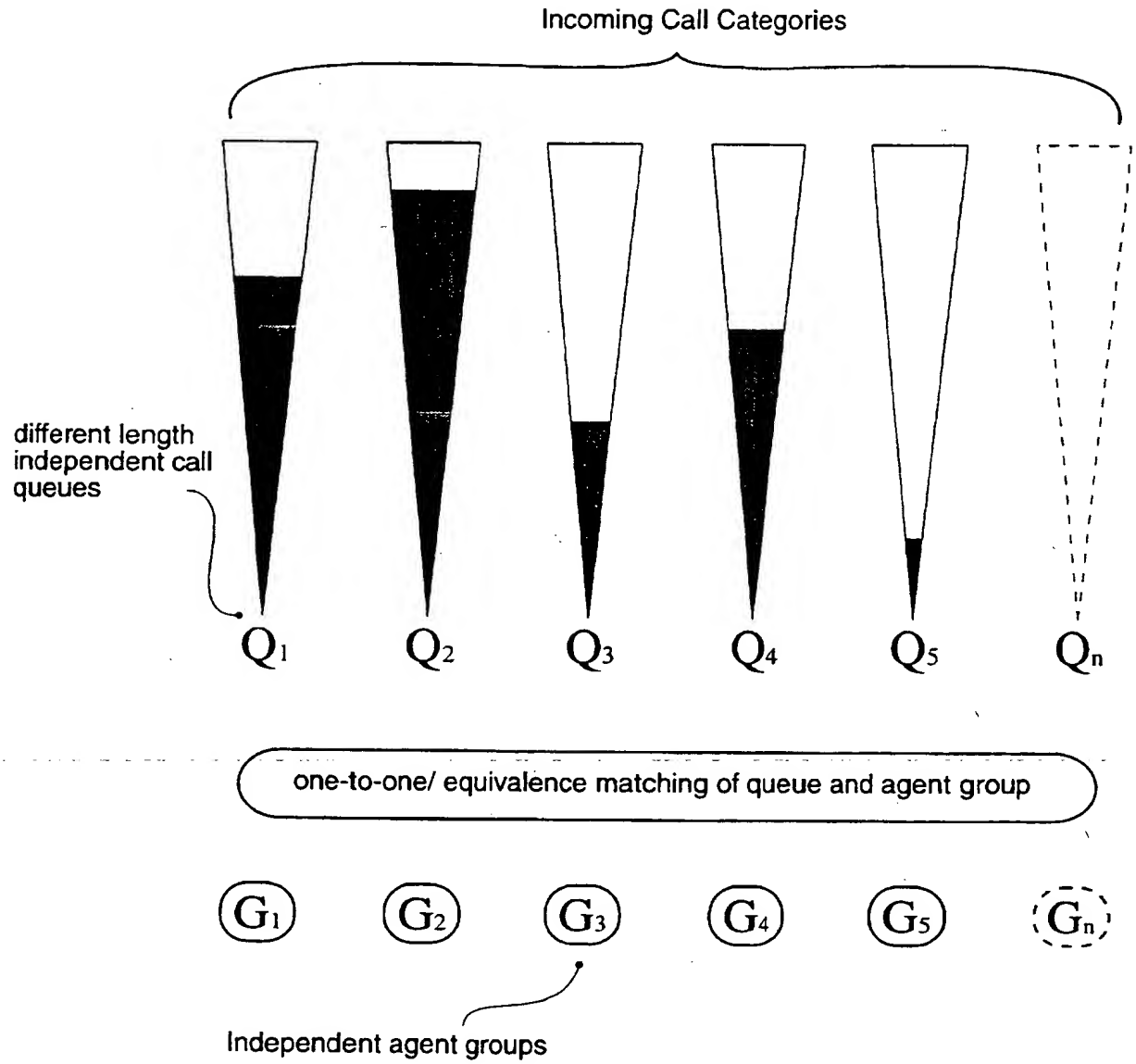

Figure 1

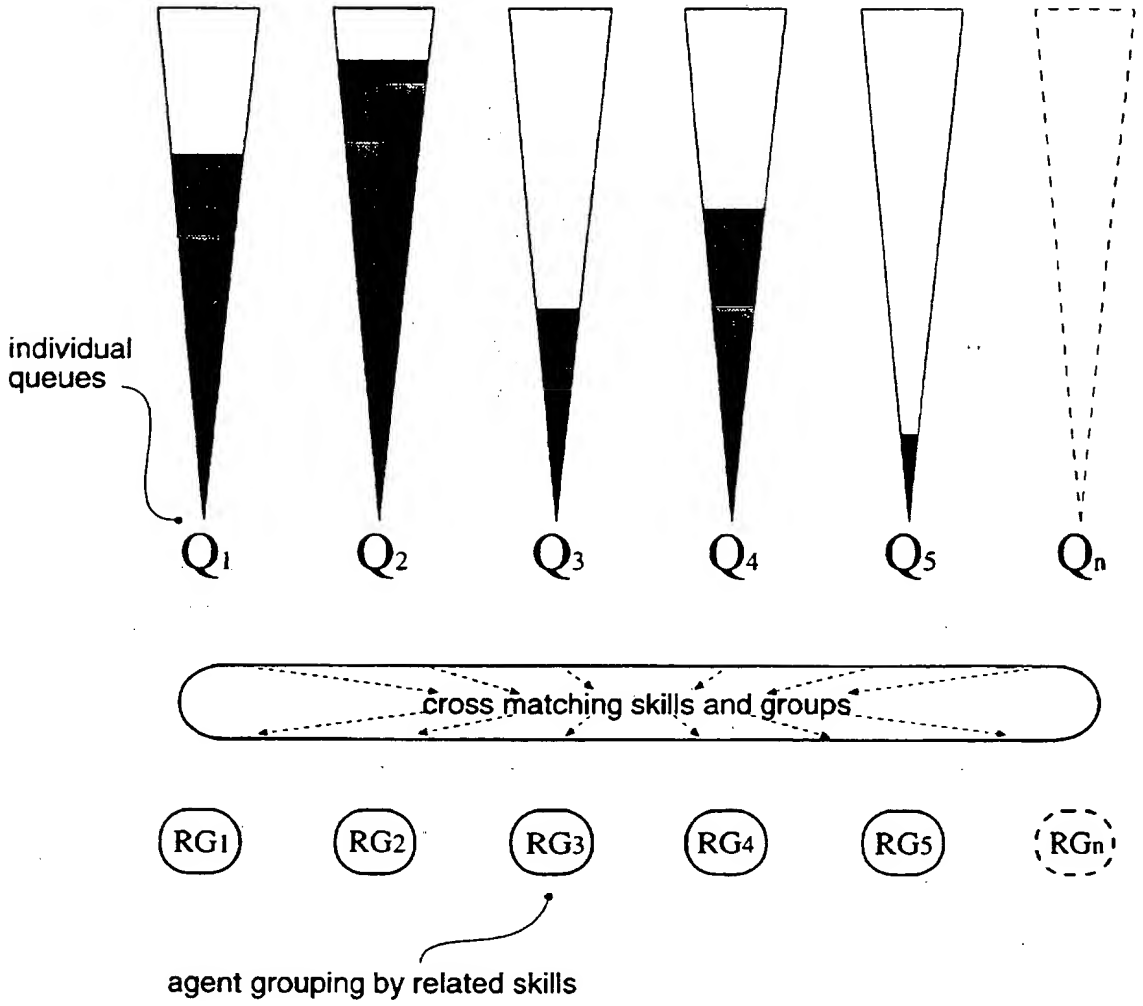

Figure 2

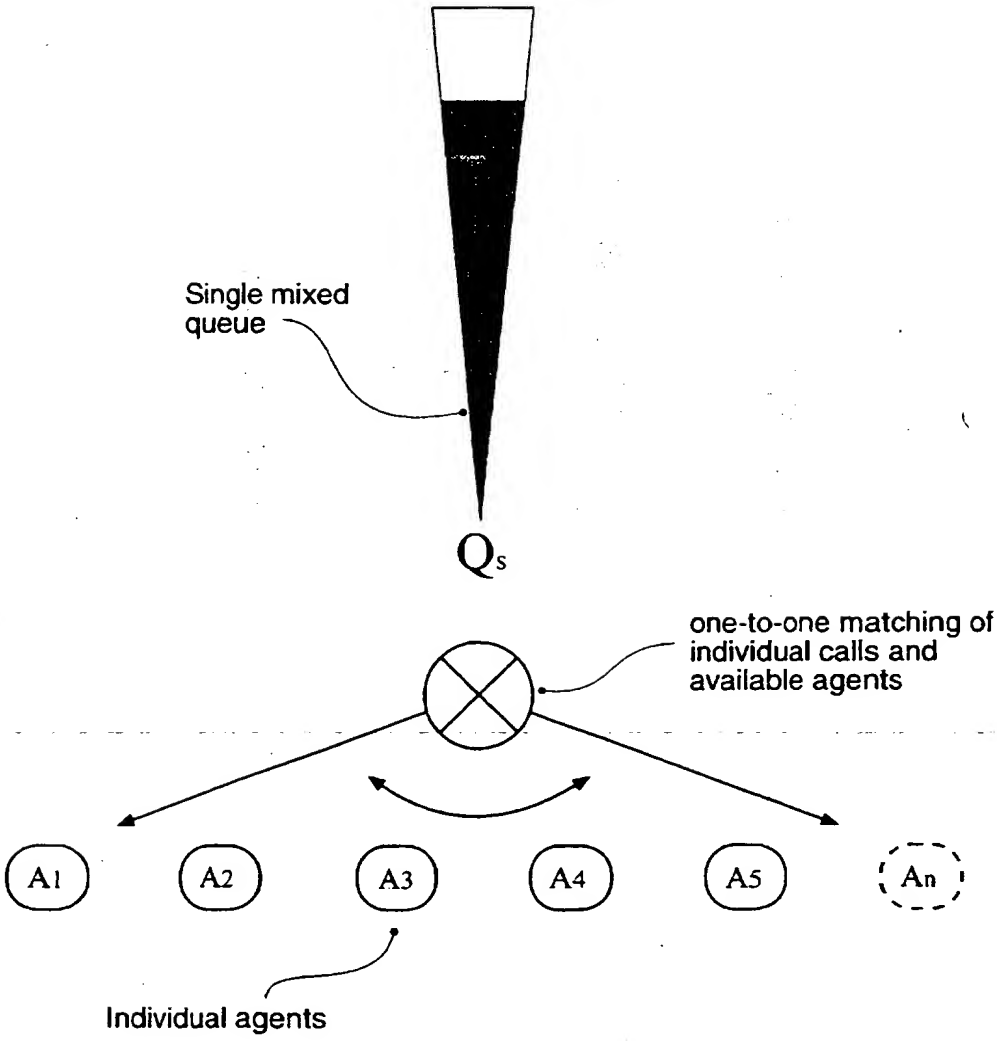
Single mixed queue
Qs
one-to-one matching of individual calls and available agents
A1
A2
A3
A4
A5
An
Individual agents

Figure 3

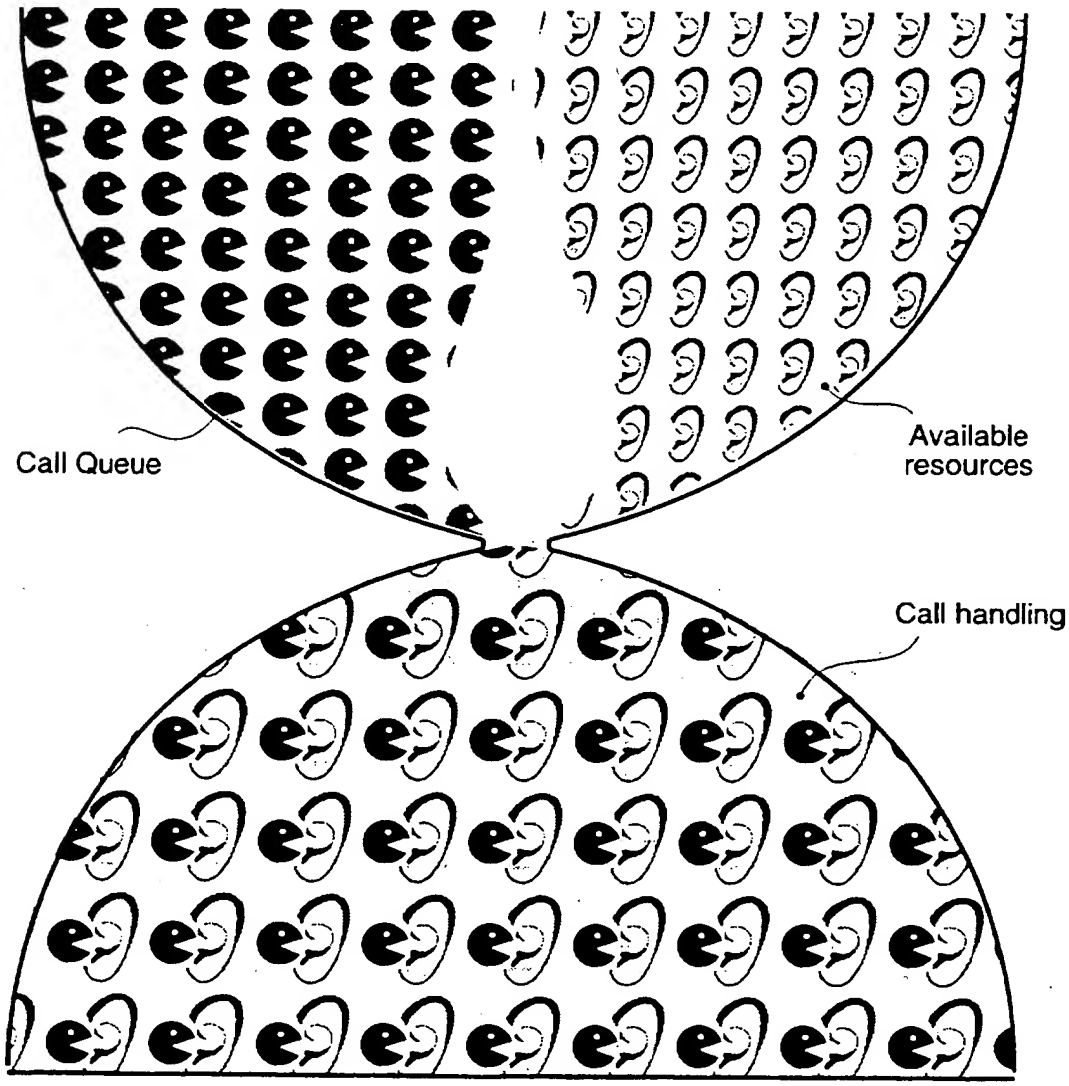

Figure 4

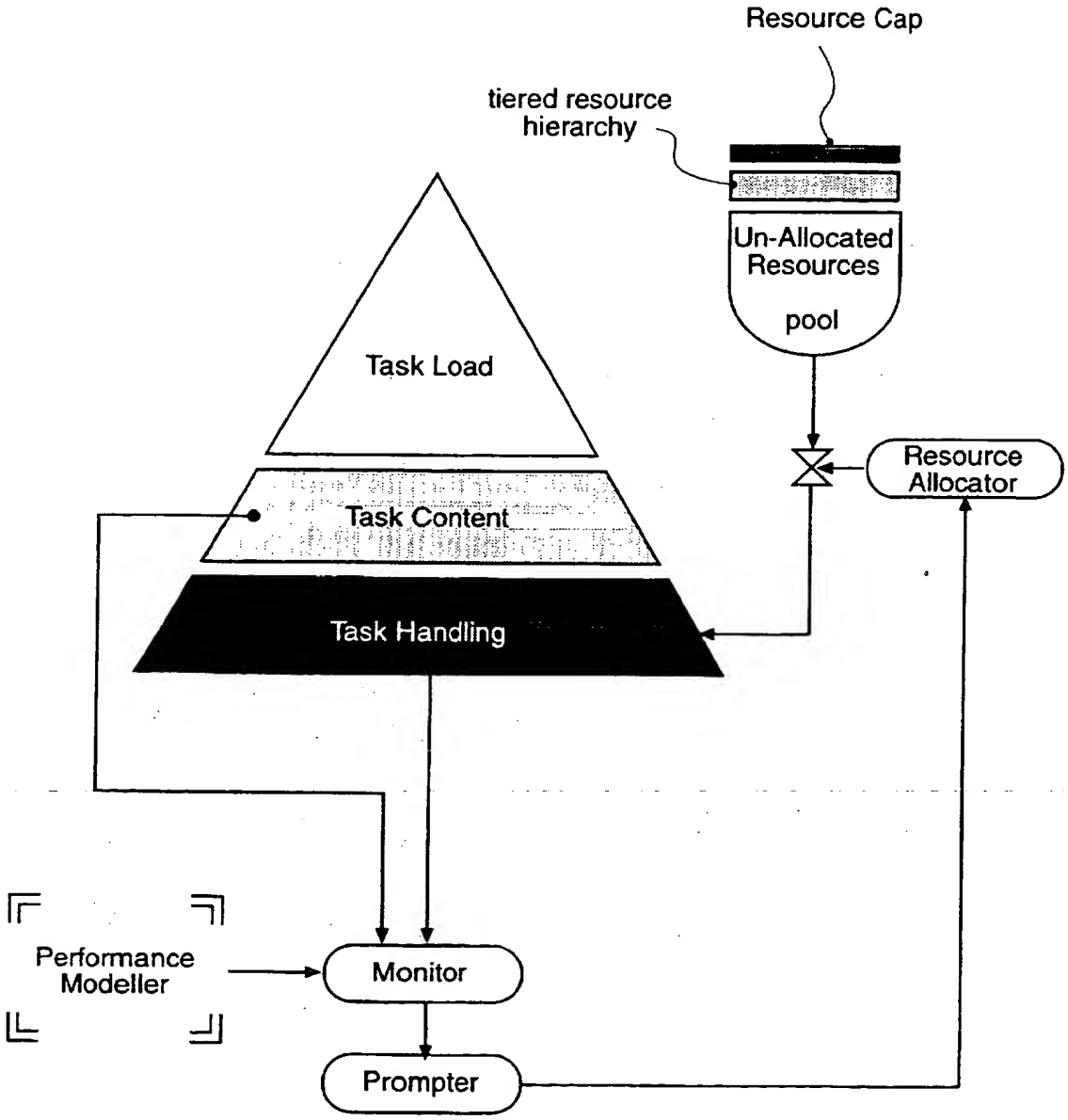

Figure 5

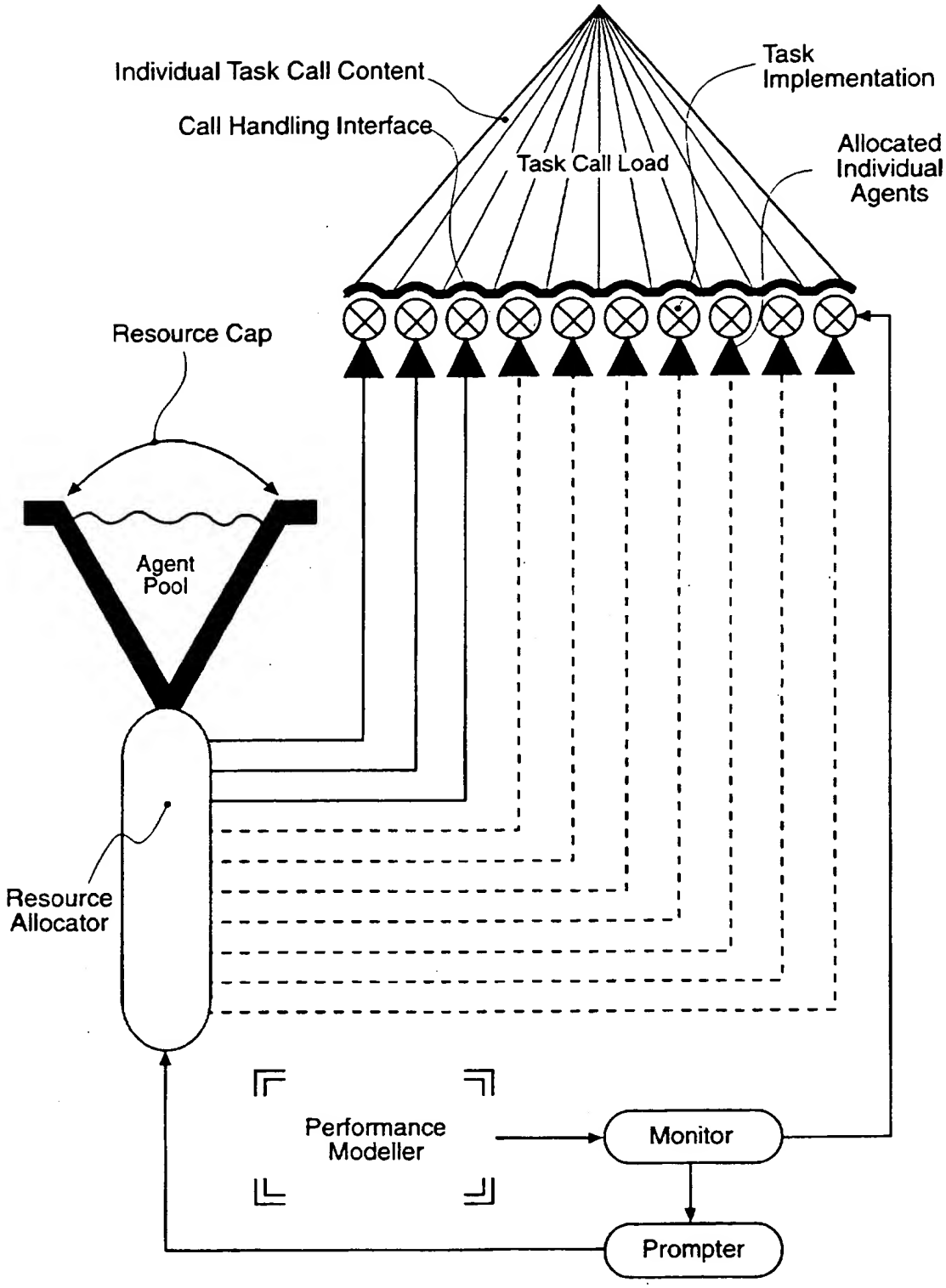

Figure 6

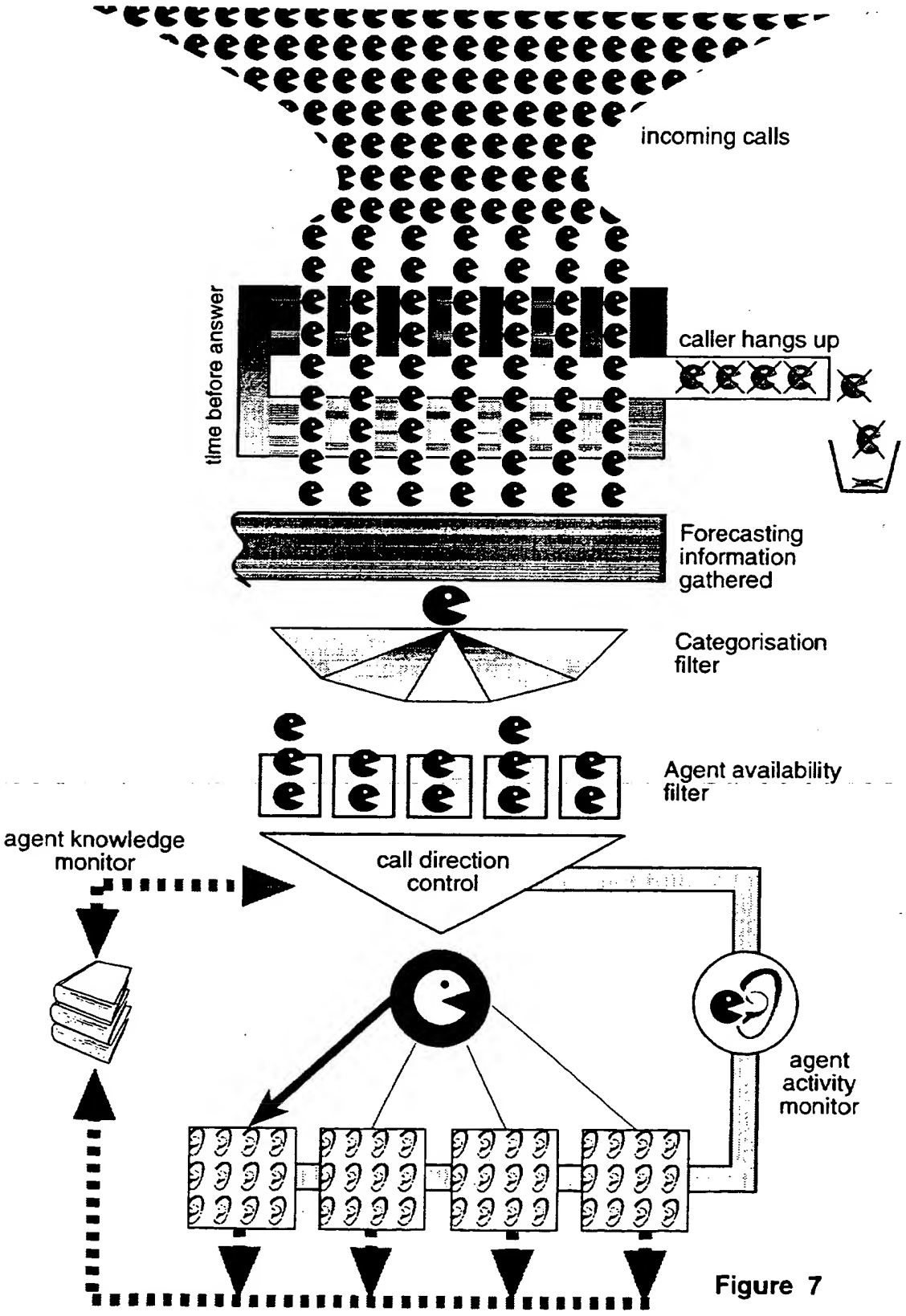

Figure 7

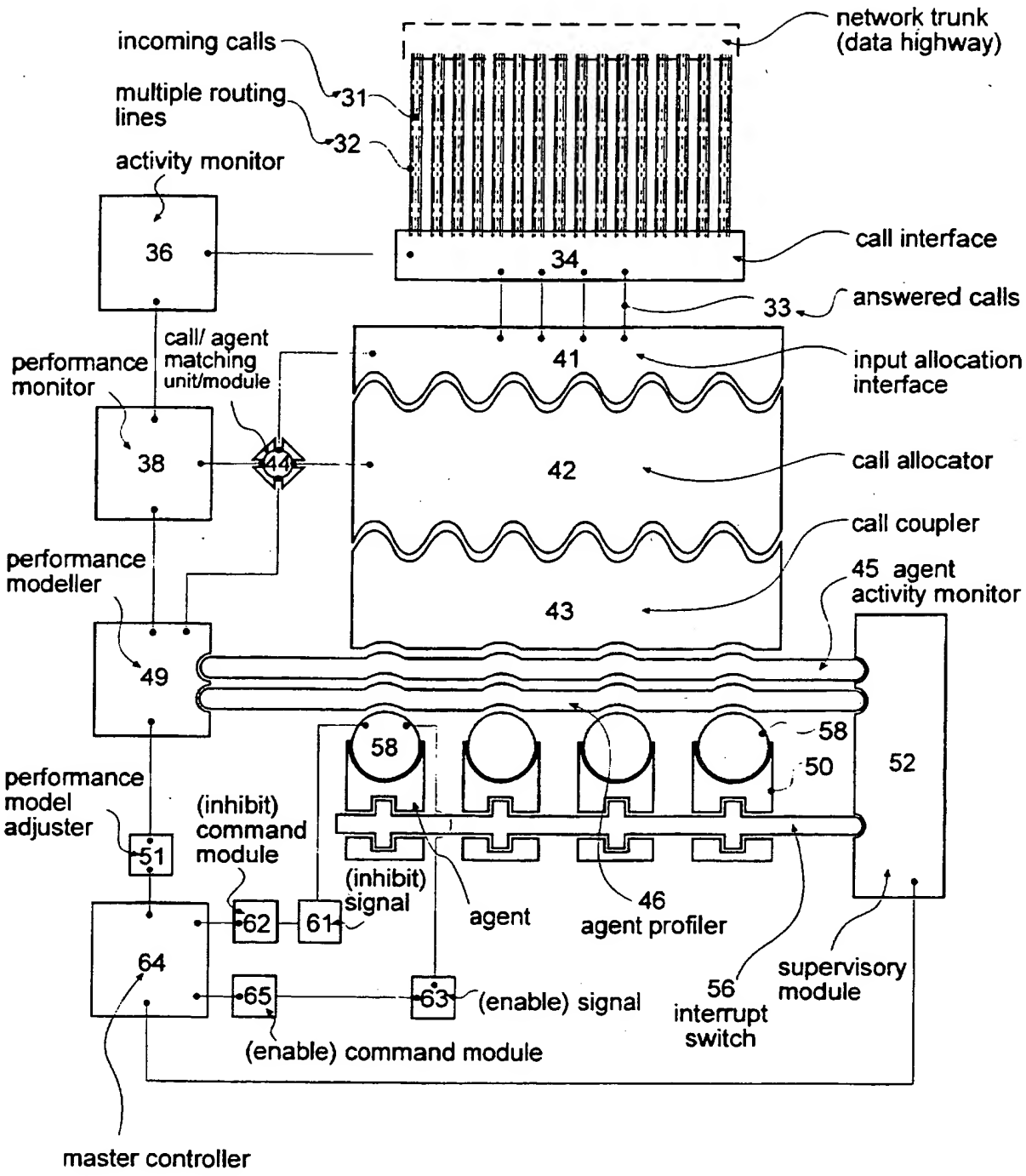

Figure 8

(Human) Resource Allocation in (Call Centre) Task Management

This invention relates to resource allocation for task undertaking, implementation or management and is particularly concerned with human resource allocation - taking account of individual and collective, grouped or pooled skills, capabilities or knowledge and availability.

The term 'task' is used herein to embrace any form of action, burden, load, action, role, exercise or job, whether stand-alone or in succession or sequence.

Similarly, the term 'resource' is used herein to embrace any form of (cap)ability, talent, potential, latency or reserve - whether human physiological, psychological, or automated, mechanised, electronic, chemical, or bio-chemical, alone or in combination.

Generally, tasks are undertaken more effectively by resources matched to task demands, burden or content - which may represent a combination of 'difficulty', variety and immediacy or timing.

That said, (capital and revenue) costs attach to resources and more 'capable' resources may well embody a higher inherent cost.

Thus, for tasks with economic constraints, in deploying resources upon a task, account must be taken of resource deployment cost.

Such cost considerations impact upon the (degree of) matching of tasks and resources.

Tasks may vary in timing or occurrence and a resource management routine must be capable of addressing variable (multiple) task load.

This adds another dimension to the resource allocation problem.

The underlying methodology of resource tasking or task resourcing admits of wide application.

Call Centre

A particular case in point is the staffing, with human agent (operators), of a telephone (voice or voice and data) call (transaction) centre.

Call 'content' may embody the provision or exchange of information or the transaction of business.

Call content is not necessarily wholly predetermined, since it may reflect an interaction with, and input from, an agent handling the call.

Thus supplementary issues, beyond a caller's original intentions, may arise, affecting call duration and 'difficulty' or span of subject-matter.

It is common practice to route calls to a call (handling) centre, dedicated to and structured for efficient call transaction.

Calls are primarily incoming, but, when capacity, or more particularly agent availability, allows, call centre agents may initiate 'cold' calling to past customers or customer prospects.

Thus, for example, an agent could follow-up or complete earlier customer-led calls, or relay special marketing initiatives or sales offers, judged relevant to that customer, from a knowledge of past transactions.

Calls may be routed from remote locations and may be switched between call centres in a network to address call traffic pattern variations.

A call centre typically incorporates multi-line, exchange switching or routing functionality and is 'tasked' with addressing incoming call traffic - of largely unknown content and variability - with the object of optimised call handling.

That said, call traffic may follow characteristic patterns, such as time of day, weather conditions, season, which when monitored as historic data may give some basis for prediction of call volume.

In evaluating a call handling routine or methodology, call traffic may be modeled to allow simulation of the performance of the call handling routine.

Various statistical models may be employed for incoming call simulation or generation, such as Stochastic, Poisson Distribution, etc.

Yet another approach uses exponential random variables, in which a continuous random variable has an exponential distribution.

A refinement could entail consideration of an hyper-exponentially distributed random variable.

Certain mathematical models or algorithms are also known for relating call traffic to agent handling - a particular case in point being the so-called Erlang B and C models.

According to one aspect of the present invention, resource allocation in telephone call demand satisfaction adopts a methodology calling upon a neural network approach to combinatorial optimisation, allowing flexible use of multi-skilled agents.

This requires fewer agents than would have been suggested by traditional statistical methods for agents with more limited skill 'bands' and calls segregated into separate queues according to content or nature to reflect those 'banded' agent skills.

Another aspect provides a more sophisticated model of incoming call distribution, using a Poisson process, or a stochastic queue based upon the Erlang C model commonly used in scheduling software.

Service (or call handling) time distribution is addressed by a hyper-exponential model, taking into account agent multi-skills.

A further refinement allows individual agent grading according to their own unique skills profile, so that incoming calls are allocated to an individual, rather than a group.

The system performance can be tested by simulation of incoming call traffic.

Optimisation of call handling need not represent an 'absolute' standard or 'perfection' in performance.

Rather, optimisation can reflect compliance with predetermined reference criteria.

This allows a degree of in-built pragmatism and compromise, reflecting real world constraints.

Thus performance criteria compliance recognises such factors as:

- minimising call queuing or queue length in terms of both:

- numbers of calls 'held' temporarily in the queue; and
- duration of holding of individual calls;

- an average 'collective' call hold or wait duration.

In principle, to a caller expecting prompt attention, any delay in call handling and substitute placement in a queue can engender disappointment, frustration or irritation.

Whilst a certain caller tolerance to delay - and even a willingness to try again - may be relied upon for a certain proportion of callers, delay promotes ill-will.

With callers as prospective customers, the nature and extent of delays merits careful scrutiny.

Moreover, in non-monopoly supply situations - such as airline seat booking - prospective customer callers may even 'retaliate', by calling competitive suppliers.

The pent-up frustration of callers held, (albeit 'temporarily') unattended, in a queue can be vented upon an (unsuspecting) agent when the call is eventually answered.

In some refinements of the invention, it is envisaged that an agent could be forewarned of the time a call has been waiting, before attendance upon it.

An agent may encounter 'queued-caller' hostility or even abuse - which is at best a distraction from the original underlying call content. This in turn requires greater agent operator skill and patience.

A frustrated, impatient caller may preface interaction with the agent allocated with a complaint - and to take longer to reach the original point or purpose of the call - and be less receptive to promotional initiatives volunteered by an agent.

Overall, there may be a lost opportunity to derive more business from a given call.

In such a 'contentious' call climate, call handling time may be inflated, thus aggravating the call queuing problem.

Agent operators can become wearied, demoralised

and so less effective, if required to deal with a stream of discontented callers, held waiting.

Nevertheless, with a held - rather than abandoned - call, at least there is an opportunity to restore customer relations when the call is (eventually) attended.

An abandoned call, particularly after a prolonged wait, may not be 'retrievable' - although early abandonment may reflect 're-call preparedness'.

Overall, abandoned calls represent immediate - and potentially long term, even permanent - lost business.

Call centres are typically employed by organisations relying upon primary customer contact by telephone for business transactions.

Typical call center operators range from financial institutions, such as banks, building societies and credit card companies through to utilities, vehicle breakdown services, mail order suppliers, holiday and travel operators and airlines.

According to one aspect of the invention, a multi-tasking resource manager comprises:

- an assessor or counter, for gauging an actual incoming task load, or volume of tasks to be implemented,
- an appraiser, for appraising, judging or determining task complexity and duration, individually and collectively in a load pattern,
- a resource activator for activating or enabling resources in accordance with task load,
- a resource allocator, for allocating resources to immediate tasks,
- a performance modeler, for modeling a desired overall performance of response to and handling of tasks,
- a monitor for monitoring the performance of tasks by different resources,
- a prompter for prompting resources, upon performance and incoming task load.

Such a resource manager may import fixed, or periodically-updated, reference models of predictive load pattern, evaluated from past actual load monitoring.

Similarly, an appraisal of past task complexity and duration may be modeled, along with resource performance, both individually and grouped.

According to another aspect of the invention a multi-tasking (agent) resource manager, for a telephone call centre, comprises

- a counter, for counting an actual incoming call load volume, of calls to be handled and answered,
- an appraiser, for judging or determining call content, nature, complexity and duration, individually and collectively in a load pattern,
- an (agent) resource activator, for activating or enabling resources in accordance with call load,
- a performance modeler, for modeling a desired overall performance of response to and handling of calls,
- an allocator, for allocating calls between agent operators,
- a monitor for monitoring the performance of calls by different agents,
- a prompter for prompting resources upon performance and incoming call load;
- an implementer for enabling resources according to anticipated call volume, complexity and duration.

In the operational management of a telephone call centre, a particular task is the allocation of (sufficient numbers of) agents (of the required skills) to provide a specific 'service level' of incoming telephone call traffic load handling.

A call centre generally offers a variety of 'services' to callers - even in the context of merely answering calls.

Usually, a given service requires attention of a particular 'skilled' or knowledgeable agent.

An objective of the call centre management is to allocate incoming calls to (most appropriate) agents, for rapid response and effective service.

In the interest of economy in funding staff deployment, a minimum number of agent resources should be [forecast] available for deployment, to offer/provide a required 'level of service'.

The 'level of service' will depend upon or reflect:

- the ability of an allocated agent to deal with a particular caller's requirement; and
- the speed within which the call is answered.

Generally, forecasting the number of agents required to provide a given 'service level' in a call centre relies upon parameters which reflect call centre 'activity' (or incoming call traffic load) in a given period (say, daily), viz:

- **call volume** - average number of incoming calls in a given time span; and
- **call duration** - average duration of a call.

Call duration may by susceptible to internal management in the call centre.

Call volume is driven largely be external factors outside call centre control.

However, management intervention may prompt a rise in call volume through, say, marketing offers.

A call centre may be forewarned to anticipate the consequential rise in demand.

In simplistic models, individual call content, complexity or character or call mix is not addressed - except for a broad categorization of calls.

Thus calls could be '(pre-)categorized' by, say, allocating identifying initial dialing codes, or post-dialing 'qualifying' codes.

Such qualifying codes are input, commonly in response to a synthesised voice prompt, after a call

to a general enquiry number.

This enables streaming, sorting or categorization into different types of enquiry.

With these parameters, it is possible to derive the **call density** - which will vary throughout the day and the week.

In a traditional telephone call centre incoming calls are sorted into different queues, corresponding to the specific type of service required.

The calls in each queue are answered only by (a group of) agents with the (same) relevant or matching skill.

In such an approach, a caller must wait until an agent with the relevant skill - 'clustered' in an agent work group of similar skills - is available.

However, upon call answering, the caller is assured of speaking to a competent agent.

Thus, in this traditional approach, the 'standard of service' will depend upon - or indeed is defined by - *the average delay before a call is answered.*

The number of abandoned calls will bear a strong correlation to this delay.

However, in reality, agents frequently have the skills to deal with more than one specific type of service.

The traditional method of dealing with incoming calls does not take advantage such multi-skilled agents.

According to one aspect of the invention a resource manager, for a telephone call centre, comprises:

- a call density monitor, for monitoring incoming call traffic activity or load, in a given time period,
- a store for storing information upon the skills or knowledge of individual agents who might be allocated to call handling, and
- an agent allocator, for allocating agents to calls, in accordance with predetermined criteria.

The approach of the present invention is to 'tag' each agent with their individual ability to answer each of the possible call types.

In one variant calls are still held in respective queues.

In another variant calls are held in a common queue.

Any call may be answered by any agent - if no agents of the highest (or most appropriate or matching), or indeed higher, skills or ability are available.

In a minimum level of service, this approach addresses call waiting times, but not the effectiveness of dealing with call content when answered.

Thus, overall, the total number (and diversity mix) of agents required in a call centre may be reduced.

A disadvantage is that calls may sometimes be answered by agents not 'best-qualified'.

This allows both under and over-qualified agent allocation.

Over-qualified allocation is material to agent resource deployment cost, rather than any loss of quality perceived by a customer caller.

In one variant agents are (notionally) grouped in 'pools' of similar skills, where each of the skill profiles would overlap other profiles.

The grouping is essentially for identification or categorisation, to enable ease of access to, (ie connection with), a particular selected agent category - rather than necessarily an individual agent.

In another aspect of the invention, described later, the 'competencies' of individual agents are addressed, along with their individual availability for each incoming call.

Reverting to the call handling management mode under consideration, when a call needs to be answered, it will be routed to one of the mixed or multi-skill (agent) 'pools'.

This routing regime is imposed, even though the allocated (agent) pool may not have 100% competency to answer that specific call.

The advantage is that the particular call does not have to await availability of an agent with guaranteed 100% (matching) competency.

Account must therefore be taken of the 'quality of the service' offered by an allocated agent.

The required number of agents will depend upon the 'standard of service' to be achieved - that is:

- the 'quality of service'; and
- the speed with which a call is answered.

Any mathematical model, operation, procedure, routine or algorithm, for computing an optimum number of agents must take into account these factors:

- the number of agents;
- the speed of answering a call; and
- the quality of the service.

Each of these may be modelled as 'cost' terms or factors - which have to be balanced to determine a minimum overall cost function.

A 'minimum cost function' will be a solution to the problem - that is the minimum number of agents required for a given 'standard of service'.

Constraints

The construction of such a cost function should reflect various constraints to ensure an acceptable solution, for example:

- the number of agents should not exceed the maximum number of calls active in a queue at any one time;
- the quality of the service should be maximised; and
- no call should be allocated to an agent with zero competency to answer it.

In order to simplify the mathematical model, these constraints may be expressed in terms of additional costs.

Such an approach could be applied to any call centre.

The skills of different groups of agents must be specified.

Once this is done, an optimisation algorithm will determine how many agents from each skills profile are required.

This methodology may be set up to operate on an 'instantaneous' basis - producing the expected number of agents required in each skills mix, for each successive (say, half-hour) operating period.

This requires a separate or self-contained optimisation for each (half-hour) period.

As this involves a non-linear optimisation process, the algorithm requires initialisation of its parameters - which are then (re-)adjusted until a minimum is found.

However, it might be that the minimum (initially) identified is not suitable.

Hence, the optimisation process may have to be run several times in succession, each time with different random start values.

This problem can be solved by using a single run in each (half-hour) period, and using a so-called 'scaled conjugate gradients' optimisation method.

In order to simplify the model, the following (tentative) assumptions may be made:

- time taken to attend to the call is constant;
- all staff cost the same;
- the number of agents with each skills mix is unlimited; and
- calls are uniformly distributed over the modelled period.

However, for more accurate results, some or all these assumptions may be replaced by parameters representing respective variables.

For example, the number of agents in the pools may be capped.

This may be applied to all agents, or only to those in highly qualified pools (such agents likely being fewer and/or more expensive).

Similarly, the cost of deploying agents may be weighted according to skills.

Various modelling methodologies are envisaged (for capping the number of agents available), vis:

- a so-called 'sigmoid function';
- a so-called 'softmax' methodology; and
- a so-called 'penalty function';

In modelling call queue characteristics, a so-called 'stochastic model' may be used to model the distribution of calls over a given (say, half-hour) period.

In an alternative approach, instead of grouping agents with similar skills profiles, agents are treated individually.

Incoming calls are then allocated to the 'best qualified' individual (immediately) available. This approach is shown in Figure 8.

However calls are 'pre-categorised', it is possible for an errant call to be mis-allocated from the outset - say by a caller responding to or identifying an incorrect or inappropriate query type, in a pre-sorting automated or pre-corded voice answering mode.

According to another aspect of the invention,
a call manager, comprises
a call interface (34),
for incoming calls (31),
along multiple routing lines (32);
an activity monitor (36);
for examining call flow through the interface (34),
to determine the difference between
the volume of arriving or incoming calls (31)
and 'departing' or routed for answer' calls (33),
and thus the size 'Q's' of the stack or queue
of unanswered or waiting calls (31)
upstream of the interface (34);
a performance monitor 38,
interrogating the activity monitor 36;
a call allocator 42,
'downstream' of the call interface (34),
charged with allocating calls (31/33) to agents (50);

an agent activity monitor (45),
whose 'attention' spans all the agents (50);
an agent profiler (46),
tracking an up-datable register of agent capabilities -
ie skills and knowledge;
a matching unit (44),
for matching call-content to agent skills/knowledge,
and directing the call allocator (42) accordingly;
an allocation interface (41),
at the input of the call allocator (42),
to extract calls (31) in turn from the queue;
a call coupler (43),
at the output of the call allocator (42),
configured to interconnect calls and agents,
according to performance matching criteria;
a performance modeller (49),
for setting performance matching criteria;
a performance modifier (51),
for adjusting targets set
for the matching unit (44) and call allocator (42);
a supervisory module (52),
for agent supervision,
enabled to access individual agents (50),
through a monitor coupling (54),
an individual agent switch (58),
prefacing or interfacing with all agents (50);
an inhibit command module (62),
directed by a master control unit (64),
to apply an inhibit signal (61),
to an individual agent switch (58);
an enable command module (65),
under direction of the master control unit (64),
to apply an enable signal (63),
to an individual agent switch (58).

According to an aspect of the invention, a scheduling method in a call centre is structured to predict the minimum number of agents required to achieve a certain service level and to provide a satisfactory quality of service, according to the expected call centre activity.

In this methodology, major unknowns are:

- the connectivities, that is the (actual or potential) interconnections, between telephone channels and agent skills groups.

The object is to find optimum values for such connectivities - ie the values that respond best to the conditions imposed.

These conditions include:

- minimum number of agents;
- a certain service level;
- a 'good' quality of service; etc.

... modelled as 'cost' terms.

This in turn leads to determination of a minimum overall cost function, reflecting the various imposed costs.

The model essentially expresses cost constraints to a cost function - effectively precluding the connectivities from taking certain values.

An algorithm is constructed to explore the scope or 'space' of possible connectivities, through a series of incremental exploratory steps, with a 'test' at each step.

The number of incoming queues of calls is denoted by factor 'Q'.

At a time (instant) 't', the call density in queue 'q' is given by $n_q(t)$.

The total number of calls offered in a given (say, half hour) period is $N_q(t)$.

The mean call duration is $r_q(t)$.

The number of pools of agents is denoted by P.

The agent skills profile in a given pools is expressed by a Q-dimensional vector s_p, given Q different types of call.

The agent workforce can be characterised with a skills matrix, of which an individual entry S_{pq} represents the ability of an agent belong to the skills mix pool p to answer a call of type q, that is from queue q.

Also considered is a 'connectivity' or connections matrix C, with entries c_{pq} denoting a link between queue q and pool p, with various constraints to preserve matrix validity.

The cost function embodies imposed constraints with variables representing connectivities - whose final values correspond to an optimum.

One such constraint is that, at any given instant, only $n_q(t)$ calls are active in the queue q.

Thus the sum of all connectivities over all pools must be less than, or equal to $n_q(t)$, ie:

$$\sum c_{pq} \leq n_q(t)$$

Another constraint is to minimise the number of agents deployed, so the sum of all connectivities should be reduced as far as possible; ie

$$\sum_{p=1}^{P} \sum_{q=1}^{Q} c_{pq}$$

should be minimised.

A final constraint is to ensure connectivities are set so that the most qualified agent available will answer a call.

Such a measure of 'service quality' is more 'complete' than merely, say, a statistical measure of the percentage of calls answered within a given period, commonly 30 seconds.

This traditional approach assumes all calls are answered by completely competent agents.

Whist call answering delay is a consideration, it is taken in conjunction with how well a call is answered.

A normalised measure of total quality of service over all queues would be:

$$\frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{n_q} \sum_{p=1}^{P} c_{pq} S_{pq}$$

This can be reflected in a quality of service or call/agent matching measure.

Minimising the cost function involves quantifying the loss in quality as:

$$E_0 = \frac{1}{Q}\sum_{q=1}^{Q}\left(\sum_{p=1}^{P} c_{pq}S_{pq} - n_q(t)\right)^2$$

This term would be equal to zero, if all calls were answered by people with 100% competencies.

Another constraint is that there are only $n_q(t)$ calls in queue q at time t. The penalty term which represents the requirement is:

$$E_1 = \frac{1}{Q}\sum_{q=1}^{Q}\left(\sum_{p=1}^{P} c_{pq} - n_q(t)\right)^2$$

The penalty term which tries to minimise the total number of staff required is:

$$E_2 = \frac{1}{Q}\sum_{q=1}^{Q}\sum_{p=1}^{P}(c_{pq})^2$$

Negative constraints are not allowed, so an explicit penalty term which forces the connectivities to take on positive values only is introduced. One such type of penalty term could be:

$$E_3 = \frac{1}{Q}\sum_{q=1}^{Q}\sum_{p=1}^{P}\frac{1}{1+\exp(-\eta c_{pq})}$$

where η is a large value which induces a strong gradient barrier around the origin.

The last constraint is that no call should be allocated to agents belonging to a pool with 0% competency to answer it.

Ultimately, the resource allocation (management) solution obtained from such methodology allows the potential agent pool to be considered as heterogenous - albeit differentiated by a set of overlapping skill mix profiles.

This (management) model would in principle allow any call to be answered by any agent in any of the skills mix pools.

The methodology returns the predicted number of agents required in each agent pool.

Further model refinements could address:

- an algorithm for allocating agents to calls - eg on a best match principle;
- temporal smoothing;
- differentiated quality of service considerations;
- alternative constraint encoding regimes;
- staff cost variability;
- limitations upon agent pools - particularly for more highly qualified agent profiles;
- non-uniform call distribution over a sampling period - and recognition of call density fluctuations;
- variability in call handling beyond average 'service' time and modelling of call (content) distribution.

A resource allocation cost function can be refined better to express agent skill variability.

Thus, if 'c' denotes the total number of agents required for a given service level and $n_q(t)$ denotes the number of agents needed to answer calls from a queue q, c is computed from, say, a stochastic model of queues and $n_q(t)$ is the fraction of the c agents expected to answer calls from the queue q.

An agent p from an agent workforce pool P can be characterised by an individual skills profile vector S_p.

With a total of Q possible skills - reflecting an ability to answer Q telephone queues - the entire workforce can be characterised by a complete skills matrix, identifying the skills profile for each agent.

An entry in this matrix S_{pq} is the ability of an agent p to answer a call type q.

With unique individual skill profiles, the overall skills matrix size is P_xQ.

A connectivity matrix C maps an array of entries c_{pq},

denoting links between a queue q and a pool p.

Connectivity can be re-appraised, with entry c_{pq} denoting a link between queue q and agent p, with the following constraints:

c_{pq} is the proportion of calls from queue q that agent p answers in a given (say, half-hour) period; so c_{pq} is less than or equal to 1.

One way of implementing this constraint is through a penalty function which adds a cost penalty term if the constraint is violated.

Thus, if c_{pq} is greater than 1:

$$E_{new} = E_{old} + \frac{\sigma}{2}(c_{pq} - 1)^2$$

The second of these constraints is that one person only can be answering one call, then the sum of all connectivities to one person must also either be zero

$$E_{new} = E_{old} + \frac{\sigma}{2}\left(\sum_{q=1}^{Q} c_{pq}\right)^2\left(\sum_{q=1}^{Q} c_{pq} - 1\right)^2$$

or one. This can be implemented by the penalty function:
A final cost function can selectively combine multiple individual cost terms, with constraints reflected in various ways.

Thus, for example, in the relationship:

$$E = E_0 + \alpha E_1 + \beta E_2$$

where α and β are Lagrange multipliers, which may be set to specific values reflecting perceived relative importance of satisfying each of the penalty terms.

There now follows a description of some specific embodiments of the invention, by way of example only, with reference to the accompanying diagrammatic and schematic drawings, in which:

Figure 1 shows an overview of a traditional call centre arrangement, with a one-to-one equivalence of calls, queued by category and agent grouping;

Figure 2 shows a call centre (call handling) management regime configured according to one aspect the present invention, admitting a limited cross-over or interchange between agent groupings and (queued) call categories;

Figure 3 shows an alternative call centre configuration to that of Figure 2, with individual call and agent matching, from a common mixed category queue;

Figure 4 shows a symbolic representation of call (task) and (agent) resource matching;

Figure 5 shows an outline hardware scheme for switching individual calls to particular agents judged appropriate to achieve prescribed performance levels;

Figure 6 shows an alternative outline hardware scheme for allocating calls between agents in accordance with modelled performance criteria;

Figure 7 shows another outline hardware scheme for directing calls to those of the available agent pool most appropriately qualified; and

Figure 8 shows a further outline hardware scheme for one-to-one call and agent matching.

Referring to the drawings, Figure 1 depicts a conventional call centre system, with an entire pool of available agents sub-divided into groups according to particular skills and incoming calls queued by category.

Queues are formed strictly according to call type and so different queues hold calls of different type.

There is no cross-over or interconnection between queues, which can therefore fluctuate in length independently.

Thus queues of quite different lengths can build up.

Figure 2 shows a call centre configured with agents grouped in skills-mix 'pools', of similar and overlapping skills profiles - again to address calls

queued by category.

In this context, 'grouping' need not necessarily require a physical association, interconnection or common boundary or access route, but rather may merely represent or reflect a 'notional or token association' for computational purposes — and a concomitant (switching) facility, to connect an individual agent to an individual call.

A call will be routed to one of the skills mix pools deemed most appropriate at a moment in time, according to agent availability.

Figure 3 shows an alternative call centre configuration, where each individual agent has a skill grade allotted for each of the various tasks which might be called upon - addressing a single mixed category call queue.

Incoming calls are allocated, on a one-to-one basis, to the 'best qualified' individual agent available, rather than to an agent pool.

Figure 4 depicts symbolically the queued call and agent resource allocation task for call to agent matching in call handling, by one of the approaches reflected in the various other Figures.

Figure 5 depicts an outline formative hardware scheme with a task handling module addressing a task load of incoming calls, reflecting task (call) content in addition to task (call) volume, in accordance with a performance modeller, to prompt allocation of a resource (agent) from a pool of unallocated available resources (agents).

Figure 6 depicts the addressing of multiple incoming calls of diverse content, through a call handling interface with multiple individual task implementation (or agent resource allocation) cells to allocate individual calls to individual agents, drawn from a common agent pool of mixed skills, in accordance with a performance modeller.

Figure 7 depicts call queuing by call category, with logging of an overspill of abandoned calls, and an information gathering module for call forecasting.

Post-arrival call categorisation follows the call queuing regime and individual agent availability is modelled through a filter, prefacing agent allocation in accordance with agent skills.

An overall agent activity monitor allows performance modelling - ie how effectively the agent resource is being deployed to meet a given incoming call demand.

Figure 8 shows an outline hardware scheme for individual call to agent matching.

Referring to Figure 8, incoming calls 31, from a network trunk 30, are distributed over multiple routing lines 32.

The individual line 32 capacity may be a single, or multiple simultaneous calls, through, say, multiplexing techniques.

The physical number of lines 32, or cumulative line capacity, is thus a constraint upon the number of calls 31 which can be handled at any given time.

Such a line capacity constraint is independent of whether there are sufficient downstream agents to handle them.

The calls are applied to a call interface 34, which prefaces an input allocation interface 41, for onward allocation of calls to agents.

An activity monitor 36 examines the call flow through the interface 34 - and so can determine the difference between the volume of arriving or incoming calls 31 and 'departing', allocated (ie to an agent) or answered calls 33, and thus the size 'Qs' of the stack or queue of unanswered or waiting calls 31.

A queue size 'Qs' thus derived is one of the performance measurement factors applied by the activity monitor 36 to a performance monitor 38.

A call allocator 42, 'downstream' of the input allocation interface 41, is charged with allocating calls to agents 50, through a call coupler 43.

Account is taken of the activity of individual agents 50, through an agent activity monitor 45, whose 'attention' spans all the agents 50, without interfering with their activity or performance.

The agent activity monitor 45 can take account of both agents 50 active or available at a point in time and agents 50 not currently active or available - but who might be turned to or 'brought into play' when circumstances dictate.

Similarly, an agent profiler 46 tracks an up-datable register of agent capabilities -ie skills and knowledge.

A call-content to agent skills/knowledge matching unit 44 bridges the input allocation interface 41, the call allocator 42, the performance monitor 38 and a performance modeller 49.

The performance modeller 49 in turn addresses the agent activity monitor 45 and agent profiler 46.

Overall, the call coupler 43 interconnects calls and agents, according to performance matching criteria set by a performance modeller 49.

Provision is made, through a performance model adjuster or modifier 51, for adjusting the targets set for the matching unit 44 and so the actions of the call allocator 42

The performance modeller 49 can also set standards for the incoming call queue size Qs.

Overall, a balance can be struck between queue size and matching.

Thus a somewhat 'looser' match between call content/category and agent skills/knowledge can be countenanced when call traffic load is high - to inhibit undue call queue length or size.

That said, gross mis-match between calls and agents can lead to longer call handling times - and thus a deterioration in capacity (ie number of agents who might be expected to be available for call handling over a given period).

This can in turn lead to an increase in queue size - that is a contrary effect to that intended by the more flexible call/agent matching regime temporary implemented.

With a 'lighter' incoming call load, a 'stricter' or closer match of call content/category and agent capability can be stipulated.

Agents can be made aware of both queue length and call/agent matching, along with time taken to handle calls.

In alternative configurations, each individual agent could be monitored through an associated switch, which determines the connection of that agent to a

call.

It is conceivable that, with some knowledge of the nature or at least broad category of each call, an individual agent could be forewarned of the conformity of that call content or character with the agent's individual capability, skills or knowledge range.

The time spent by an agent on a particular call could then be monitored in the context of the match between call and agent.

Thus a gross measure of agent performance by call handling time could be replace by a more sophisticated measure of call handling time offset by call mis-match.

In a further refinement, an agent could be deemed to be in learning or training mode - with a concession on call handling time expectation for calls with a 'low' match factor.

An agent would thus not be discouraged by a low call handling rate.

Provision is made for agent supervision through a supervisory module 52, which is enabled to access individual agents through a monitor switch 54.

Supervisor monitoring, or active intervention or participation, in agent handling of a call could also be taken into account in the performance modeller 49 and performance modifier 51.

Routing of subsequent calls could reflect the pattern of call to agent matching of preceding calls.

Thus an agent could be relieved of the burden of repeated mismatch, or a deliberate degree of mismatch allowed or even contrived to extend and develop agent skills.

This 'human factors' element could also be implemented through the performance modeller 49 and performance modifier 51.

In some configurations, an individual agent switch 58 could be instituted as part of a multiple switch, prefacing or interfacing with all agents.

An inhibit signal 61 could be applied to an individual switch from an inhibit command module 62,

responsive to overall direction from a master controller 64 and/or the performance modeller 49.

Similarly, an enable signal 63 could be applied, from an enable command module 65, under the direction of the master controller 64 and/or the performance modeller 49, to an individual switch 58, to allow the associated agent 50 to participate in call handling.

The agent activity monitor 45 addresses the individual agent switches, to sense occupation by an agent 50 with a call.

An interrupt switch line 56 from the supervisory module can enable or disable agent involvement, either before or during a call.

Similarly, through such an interrupt line 56 a supervisor can intervene in individual agent call handling activity, for performance monitoring, supervision, guidance and training.

Moreover, an agent could seek guidance from the supervisor by voluntarily using the interrupt line 56 to address a supervisor through the supervisory module 52, or even another agent, designated as a mentor.

This in turn could be used to ensure that, once an agent is occupied upon a particular call, no further calls can be applied to it.

Passive data collection upon such basic factors as individual call duration - or the time taken by an individual agent to handle a call - could be accumulated alongside active call switching to unoccupied agents in a 'priority hierarchy' - factoring fitness or match considerations - determined by the control unit.

In the multi-channel highway 30 used to route incoming calls to agents, signal encoding could be employed to keep calls separate and to authorise an intercept by the next allocated agent, provided with the relative signal code an enabling authority.

In principle, this would allow more than one agent to share a call - for example under agent training and supervision.

Generally, provision is made, through intercept or interrupt 56. for a supervisor to monitor any agent - and to interrupt or intervene as the supervisor may judge appropriate.

Similarly, the supervisor, using the facilities of the supervisory module 52, could over-ride agent allocation directions from the control unit - or adjust the deemed agent skills, through an upgrade or downgrade according to perceived and measured performance.

Information upon the overall queue length is commonly made available to agents in a call centre through a visual display (not shown) - as an agent group performance monitor and incentive.

A similar measure may be applied, even when agents are not strictly grouped, or there is a flexible (ie expansive) group definition according to call traffic load.

Such a display could also signal this group re-definition in an 'all hands to the pump' mode, which might offer even more incentive to rapid call turn-around.

A supervisor could monitor the effectiveness of the activity and progress display as an incentive - to ensure that the task presented seems feasible and not overly burdensome.

A parallel approach could be adopted in training and developing agent skills.

Thus a supervisor could progressively introduce an agent to a wider skills grouping, so testing their flexibility and breadth in coping with diverse call content.

This would enable a manual update of the resources pool to which the control unit could apply its automated allocation decision criteria.

In a particular memory configuration, the knowledge base of agent skills could be sub-divided into sections allocated to each agent.

An agent availability register could also be sub-divided by individual agent.

Memory access would be configured so that the agent availability register would be consulted first, before impacting upon the memory resource.

Thus an update upon an individual agent skills could readily be targeted to the relevant memory address.

A memory map is provided to present an overall picture of the multiple skills or resource base, by interrogating individual skills memory locations.

Similarly, an availability map is provided to link co-operatively with the availability register.

Management information upon overall skills and availability is combined in the master controller 64.

Provision may be made for tracking the activity of individual agents when engaged on tasks other than call handling.

Such tasks could be categorised according to relative importance to call handling and the agent automatically brought back into availability for call handling if found to be engaged on less important (non-call handling) tasks.

Similarly, call categories could be accorded a weighting or value, which could be balanced with the most valuable other (non-call handling task) to which that agent might turn, either upon the agent's own initiative, or following resource allocation instructions.

The performance modeller 49, for relating call traffic to agent deployment, may follow one of several alternative algorithms - such as the so-called Erlang models B or C.

The modelling objective is broadly to determine the number of agents required to provide a given quality of service and speed of call answering.

A constraint generator exports limitations upon the modeller - such as prohibiting call answering by wholly unqualified agents, maximising quality of service (defined in a particular way) and limiting the number of agents to less than the maximum number of calls queued, whether queued in categories or collectively.

The performance modeller 49 also embodies modelling methodology for capping the number of agents made available for call handling, according to mathematical functions, such as the 'sigmoid function' the 'softmax method' and 'penalty function'.

The performance modeller can also embody a call traffic monitor, for determining call density over a predetermined time period; a store for information upon agent skills; and an agent allocator for

matching individual calls and agents or agent groups.

An incoming call traffic modeller may also be provided, to evoke call traffic variability in given time periods.

Such a traffic modeller can reflect stochastic, Poisson or other randomised distributions.

An optimiser may be deployed for repeated modelling runs with different, randomised, start conditions.

An assumption unit can be deployed to import assumptions such as call attendance time, staff costs, agent totals in skills categories and call distribution over time.

Modelling can thus be undertaken with raw assumptions for ease of modelling and attendant computation - and then progressively more refined assumptions can be imported for more sophisticated modelling and laborious and complex computation.

An optimiser enables repeated computation runs until a computational minima is identified, representing a minimum cost.

More specifically, the minimum number of agents required for a given standard of service.

Claims

1.

A (call centre) telephone call (handling) manager, including:

- a knowledge base, of individual agent operator (call-handling) skills,
- a call activity monitor, for determining call density, by monitoring incoming call traffic, over a prescribed period,
- a performance reference model, reflecting target call-handling response, (adjusted by resource availability and operational budget cost constraints),
- to provide an indication of the (minimum) number of agents required to achieve a given level of service.

2.

A call manager, comprising
a switch for individual agents
charged with handling calls;
an enable line for each switch,
to authorise call switching for that agent;
an inhibit line for each switch,
to suppress use of that agent in call handling;
a command line for effecting call switching to the associated agent;
a control unit for issuing switch commands,
to join an agent to the overall call handling activity;
a store for accumulating knowledge
of individual agent skills;
a store for a performance reference model in call handling;
a monitor for determining call handling activity;
a comparator, for comparing actual call handling
activity with the reference model;
a command over-ride for using comparator output to issue switching commands;
to include or exclude agents
from active call handling.

3.

A call handler comprising:

a plurality of lines (32);
a call interface (34);
a call allocator (42);
a call coupler (43);
an agent activity monitor (45);
an agent profiler (46);
an (interface) activity monitor (36);
a performance monitor (38);
a performance modeller 49;
a performance modifier (51);
a master controller (64);
a supervisory module (52);
an individual agent switch (58);
an interrupt switch (56);
an inhibit command module (62);
an enable command module (65);
selectively deployable,
under predetermined performance criteria,
for allocating calls to agents.

4.

A call handler, comprising
a call interface (34),
for incoming calls (31),
along multiple routing lines (32);
an activity monitor (36);
for examining call flow through the interface (34),
to determine the difference between
the volume of arriving or incoming calls (31)
and 'departing' or routed for answer' calls (33),
and thus the size 'Q's' of the stack or queue
of unanswered or waiting calls (31)
upstream of the interface (34);
a performance monitor 38,
interrogating the activity monitor 36;
a call allocator 42,
'downstream' of the call interface (34),
charged with allocating calls (31/33) to agents (50);
an agent activity monitor (45),
whose 'attention' spans all the agents (50);
an agent profiler (46),
tracking an up-datable register of agent capabilities -
ie skills and knowledge;
a matching unit (44),
for matching call-content to agent skills/knowledge,

and directing the call allocator (42) accordingly;
an allocation interface (41),
at the input of the call allocator (42),
to extract calls (31) in turn from the queue;
a call coupler (43),
at the output of the call allocator (42),
configured to interconnect calls and agents,
according to performance matching criteria;
a performance modeller (49),
for setting performance matching criteria;
a performance modifier (51),
for adjusting targets set
for the matching unit (44) and call allocator (42);
a supervisory module (52),
for agent supervision,
enabled to access individual agents (50),
through a monitor coupling (54),
an individual agent switch (58),
prefacing or interfacing with all agents (50);
an inhibit command module (62),
directed by a master control unit (64),
to apply an inhibit signal (61),
to an individual agent switch (58);
an enable command module (65),
under direction of the master control unit (64),
to apply an enable signal (63),
to an individual agent switch (58).

5.

A call handling manager,
substantially as hereinbefore described,
with reference to and as shown in,
Figures 2 through 8 of the accompanying drawings.

Application No: GB 9810573.7 **Examiner:** Al Strayton
Claims searched: 1 **Date of search:** 23 August 1999

Patents Act 1977
Amended Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.Q): H4K: KF50A; KF50E; KF50X

Int Cl (Ed.6): H04M

Other:

Documents considered to be relevant:

Category	Identity of document and relevant passage		Relevant to claims
A	GB 2 315 384 A	(MITEL)	
X,Y	WO 96/22650 A1	(BT) See esp:p.74, l.29 - p.76, l.28; fig.58A, step 5805	1
Y	WO 92/07318 A1	(IEX) See the abstract	1
Y	US 5 335 268	(KELLY...) See the abstract	1

31

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.